Análise de Regressão

Mariana Costa Freitas

2024-05-03

Introdução

A atividade pesqueira desempenha um papel crucial na economia global, fornecendo alimento, sustento e meios de subsistência para milhões de pessoas em todo o mundo. No entanto, a produtividade das pescarias é fortemente influenciada por uma série de fatores complexos, que vão desde condições ambientais até fenômenos climáticos globais.

Este estudo se concentra na análise da relação entre a captura por unidade de esforço (CPUE) e uma variedade de variáveis que afetam a atividade pesqueira. A CPUE, um indicador fundamental da eficácia da pesca, representa a quantidade de peixes capturados por unidade de esforço de pesca, como o número de dias de pesca.

Entre os fatores considerados, estão a frota de pesca, o ano e trimestre da pesca, as coordenadas geográficas (longitude e latitude) e a presença dos fenômenos climáticos El Niño e La Niña durante o período de pesca.

A frota de pesca pode variar em tamanho, capacidade tecnológica e métodos de pesca, influenciando diretamente na eficiência da captura. O ano e trimestre da pesca refletem possíveis tendências temporais na atividade pesqueira, enquanto as coordenadas geográficas indicam a localização específica das operações de pesca, levando em conta a influência das características do habitat marinho.

Além disso, os fenômenos climáticos El Niño e La Niña exercem um impacto significativo nas condições oceânicas e atmosféricas, afetando a distribuição e abundância de espécies marinhas. A relação entre esses fenômenos e a captura de peixes é complexa e multifacetada, envolvendo alterações na temperatura da água, padrões de correntes oceânicas e disponibilidade de alimento para os peixes.

Para entender melhor como esses diversos fatores influenciam a captura por unidade de esforço, recorremos à análise de regressão estatística. A análise de regressão permite identificar e quantificar as relações entre uma variável dependente (no nosso caso, a CPUE) e uma ou mais variáveis independentes (frota, ano, trimestre, coordenadas geográficas, El Niño e La Niña).

Análise dos Dados

Neste trabalho estamos utilizando a base de dados “pesca”, na qual foi adicionada a informação se houve ou não os fenômenos naturais *El Niño* ou *La Niña*, obtida a partir das variações de temperatura informadas pelo Climate Prediction Center do National Weather Service. A seguir descrevemos as variáveis utilizadas para as análises:

- frota: frota da pesca, podendo ser Santos ou Ubatuba
- ano: ano da pesca, que abrange os anos de 1995 a 1999
- trimestre: em que trimestre do ano ocorreu a pesca
- latitude: latitude em que ocorreu a pesca
- longitude: longitude em que ocorreu a pesca

- fenomeno: qual fenômeno natural ocorreu, podendo ser “nino”, “nina” ou “neutro” nos casos de *El Niño*, *La Niña* ou não ocorrência de nenhum dos dois, respectivamente
- cpue: captura por unidade de pesca, ou seja, quantidade de peixes batata capturados em kg, dividida pelo número de dias de pesca

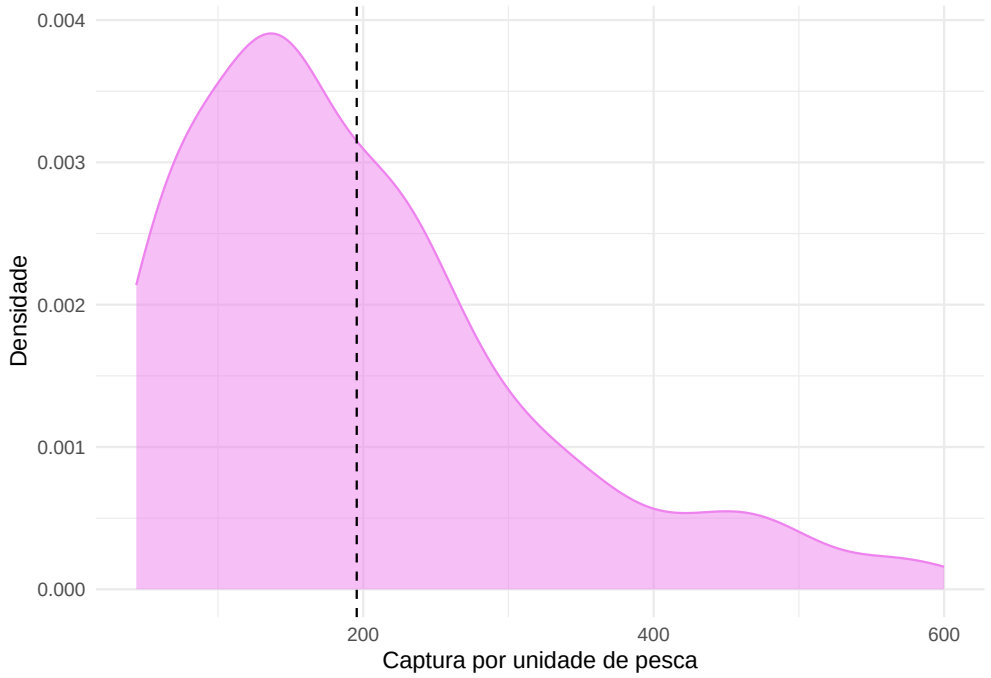
Análise Descritiva

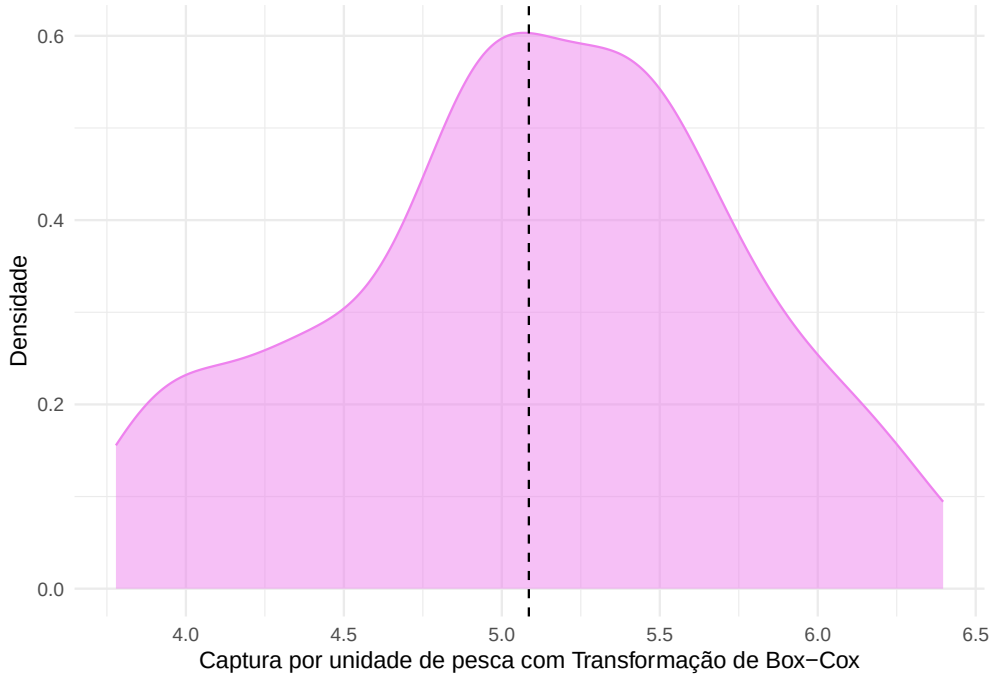
Para melhor compreensão das variáveis a serem analisadas, vamos observar algumas de suas medidas descritivas:

	N	Mean	SD	Min	Q1	Median	Q3	Max
ano	156	1997.55	1.46	1995.00	1996.00	1998.00	1999.00	1999.00
trimestre	156	2.68	1.07	1.00	2.00	3.00	4.00	4.00
latitude	156	26.22	1.24	23.25	25.25	26.25	27.25	28.25
longitude	156	46.28	1.07	41.25	46.25	46.25	46.75	48.25
cpue	156	195.55	121.06	43.75	105.56	166.41	250.00	600.00
cpue2	156	5.09	0.64	3.78	4.66	5.11	5.52	6.40

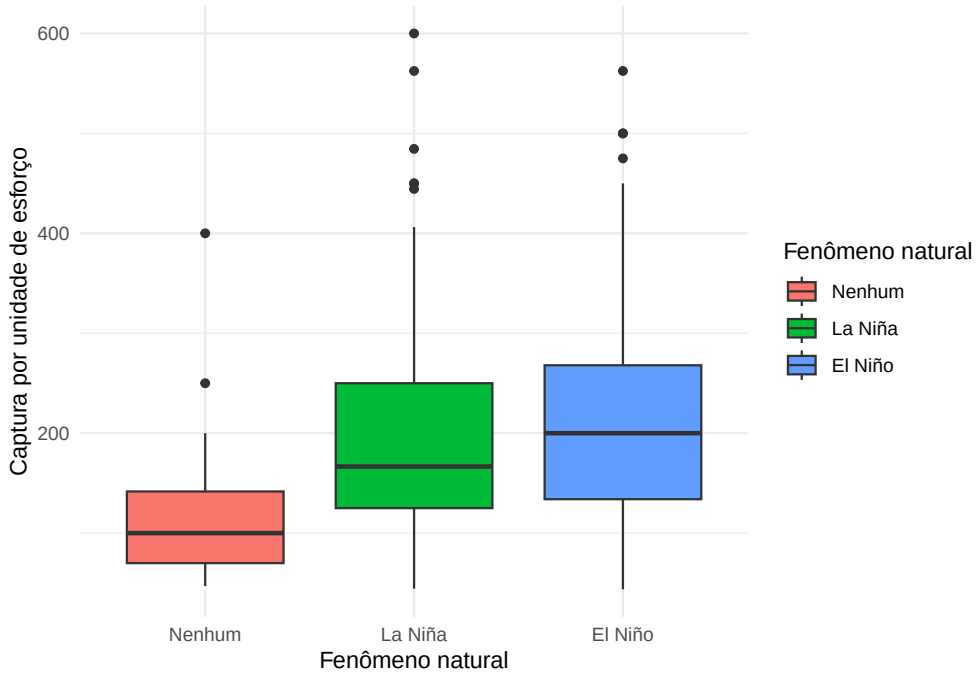
	Level	N	%
frota	Santos	117	75.0
	Ubatuba	39	25.0
fenomeno	neutro	21	13.5
	nina	93	59.6
	nino	42	26.9

Também é importante verificar como a variável que vai ser modelada se distribui. Originalmente, temos a variável `cpue`, captura por unidade de esforço. Porém, como observado abaixo, essa variável apresenta uma assimetria à direita, assim usamos a Transformação de Box-cox com $\lambda = 0$ para tornar os dados mais simétricos e semelhantes à distribuição normal. Abaixo podemos observar a diferença na distribuição da variável `cpue` (variável original) e `cpue2`, (variável com Transformação de Box-Cox).

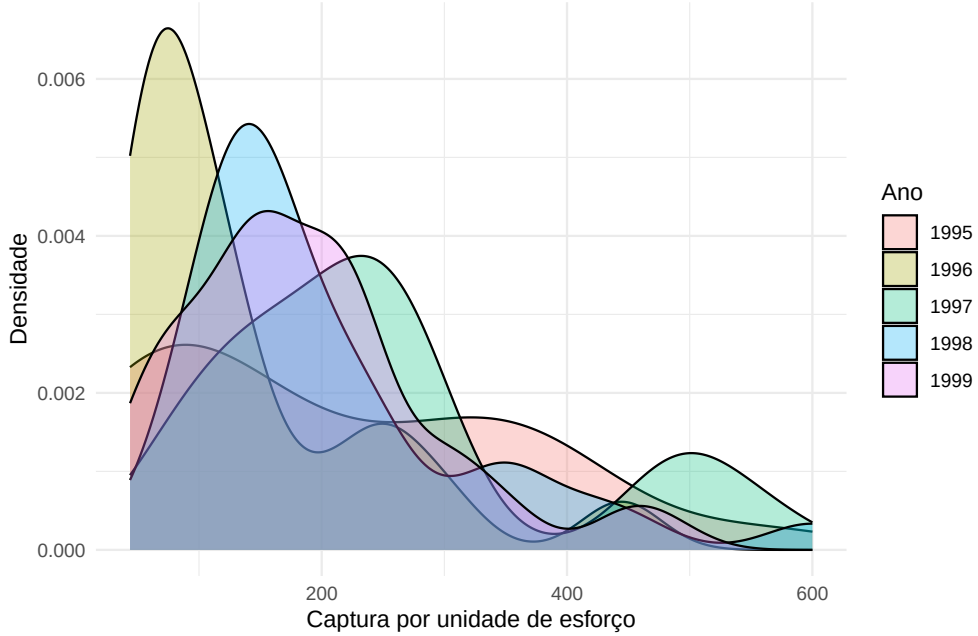




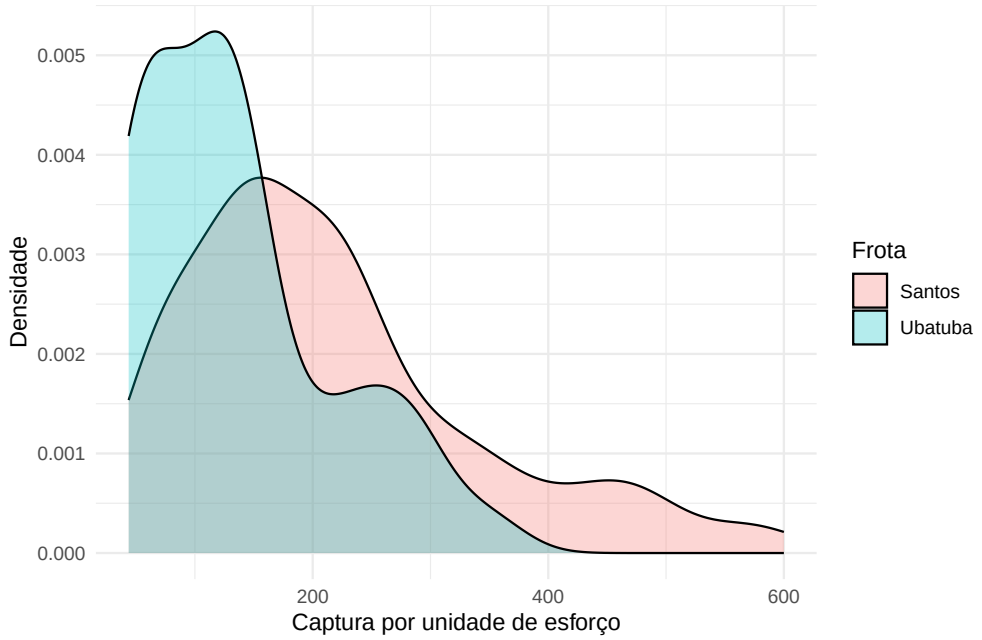
A seguir vamos analisar como o comportamento da variável `cpue`é influenciada por outras variáveis.



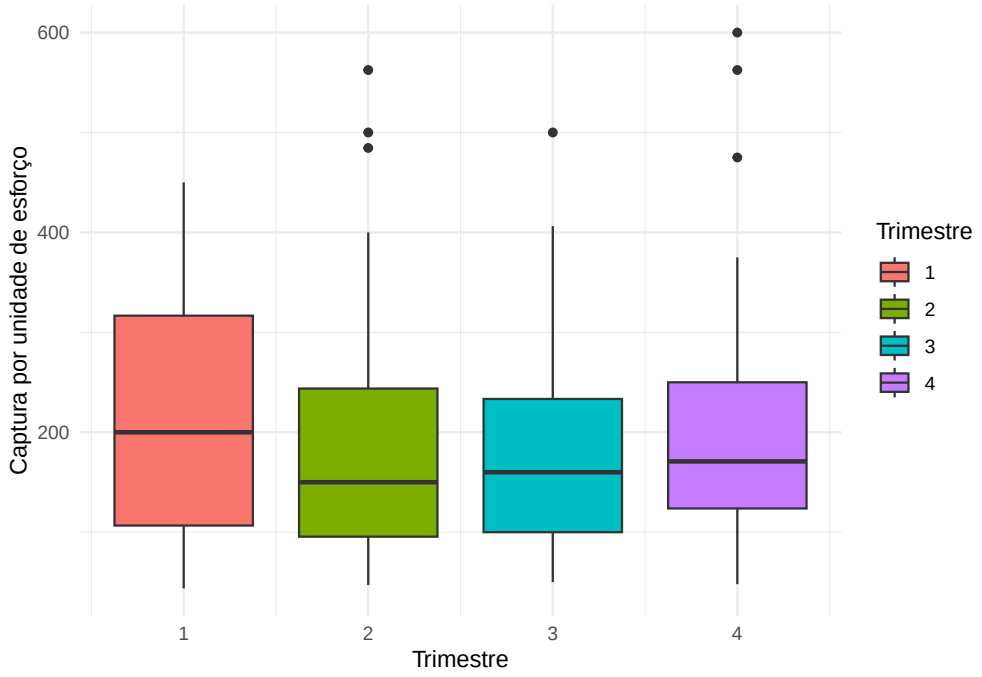
No gráfico acima podemos observar os efeitos dos fenômenos naturais *El Niño* e *La Niña* na captura de peixes por unidade de esforço. Notamos que durante o fenômeno *La Niña*, caracterizado pelo resfriamento das águas do Pacífico, e do *El Niño*, que provova um aumento na temperatura das águas do pacífico, há um significativo aumento na captura quando comparado à ausência de fenômenos naturais, com variabilidade também maior.



Nesse caso, estamos analisando como a captura por unidade de esforço se distribui por ano. Observamos que em 1995 essa variável se distribuía de forma mais homogênea, enquanto nos anos seguintes houve grande concentração em torno de 200 kg de peixe por dia de pesca.



No gráfico acima, estamos interessados em observar a distribuição da captura por unidade de esforço por frota, que pode ser Santos ou Ubatuba. É possível observar que em Ubatuba a captura se concentra em valores menores, enquanto em Santos se concentra em um pouco acima e atinge valores mais altos.



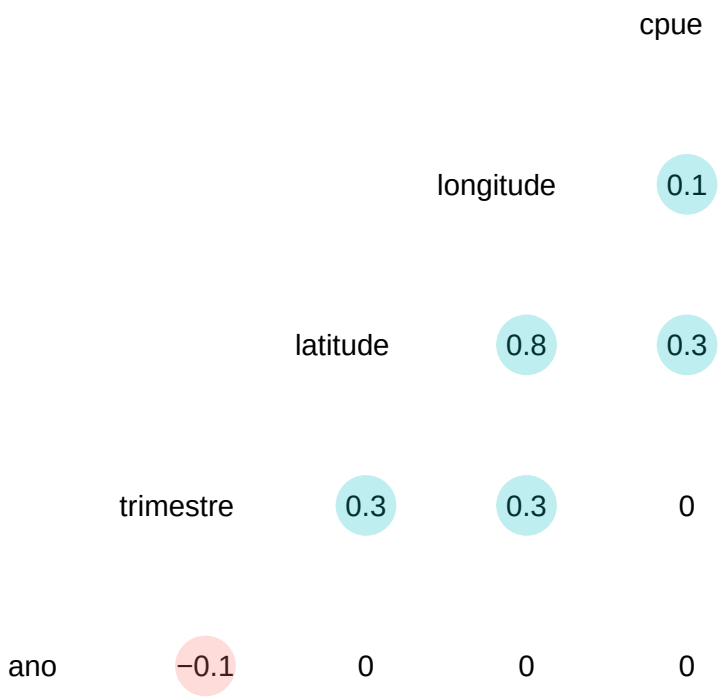
Aqui analisamos as medidas dos quartis e a variabilidade da captura em cada trimestre do ano. Podemos notar que no primeiro trimestre a mediana da captura por unidade de edforço é um pouco maior que a dos outros semestres, apresentando também maior variabilidade. Já os semestres seguintes não apresentam grande diferença nesses medidas entre si.

Correlação

A correlação descreve como as mudanças em uma variável estão associadas às mudanças em outra variável. A correlação é expressa por um coeficiente de correlação, que varia de -1 a 1. Um coeficiente de correlação próximo de 1 indica uma forte correlação positiva, o que significa que as duas variáveis tendem a aumentar juntas. Um coeficiente de correlação próximo de -1 indica uma forte correlação negativa, onde uma variável tende a diminuir quando a outra aumenta. Um coeficiente de correlação próximo de 0 indica que não há correlação linear entre as variáveis.

A correlação é importante na análise de regressão porque ajuda a entender a relação entre as variáveis independentes e a variável dependente. Antes de construir um modelo de regressão, é crucial examinar a correlação entre as variáveis independentes e a variável dependente. Se houver uma correlação forte entre uma variável independente e a variável dependente, isso sugere que a variável independente pode ser um bom preditor da variável dependente e pode ser incluída no modelo de regressão.

A seguir é possível visualizar a correlação entre as variáveis numéricas que estamos trabalhando:



Nota-se que, com exceção das variáveis `latitude` e `longitude`, que apresentam alta correlação, as variáveis apresentam baixa correlação ou nenhuma correlação entre si.

Modelagem

Um modelo de regressão linear visa descrever a relação entre uma variável dependente (também chamada de variável de resposta) e uma ou mais variáveis independentes (também conhecidas como preditoras ou explicativas). Para selecionar o modelo que melhor se ajusta aos dados, vamos usar como base o Critério de Informação de Akaike (AIC), que é uma medida que apresenta menor valor para o melhor modelo. Logo, nosso objetivo é encontrar o modelo que apresenta o menor AIC.

Para isso, vamos utilizar o método de seleção de variáveis chamado *Stepwise*, que começa com o modelo completo, ou seja, com todas as variáveis no modelo e remove ou adiciona as variáveis caso uma dessas opções gere uma diminuição no AIC. Quando nenhuma dessas ações ocasiona um menor AIC, consideramos que o melhor modelo foi obtido.

Em R, executamos esse processo primeiro utilizando a função `lm()`, que ajusta o modelo aos dados, estimando os coeficientes que melhor ajustam os dados observados. Em seguida, usamos a função, `stepAIC()`, com argumento `direction="both"`, que funciona seguindo um procedimento iterativo que envolve adicionar ou remover variáveis independentes do modelo, uma de cada vez, e comparar os valores do AIC para determinar se a adição ou remoção da variável resulta em uma melhoria no ajuste do modelo. O processo continua até que nenhuma alteração adicional resulte em uma redução significativa no AIC.

Ao executar esse processo aplicado aos dados de pesca no R, obtemos que as variáveis presentes no modelo para descrever a variável `cpue2` são frota, latitude, longitude e fenômeno, com coeficientes -0.20244, 0.19974, -0.10525, 0.33883 (em caso de *La Niña*) e 0.47914 (em caso de *El Niño*, respectivamente).

Conclusão

Neste estudo, exploramos a relação entre a captura por unidade de esforço (CPUE) e uma série de variáveis que influenciam a atividade pesqueira, incluindo frota, latitude, longitude e a presença dos fenômenos climáticos El Niño e La Niña. Utilizando análise de regressão estatística, desenvolvemos um modelo que incorpora essas variáveis para descrever a CPUE.

Os resultados obtidos demonstram que a frota de pesca desempenha um papel significativo na determinação da captura por unidade de esforço, refletindo as diferenças nas capacidades técnicas, métodos de pesca e estratégias de operação entre as embarcações. Além disso, as coordenadas geográficas (latitude e longitude) mostraram-se importantes para capturar variações na distribuição e abundância de recursos pesqueiros em diferentes regiões.

A inclusão dos fenômenos El Niño e La Niña no modelo revelou sua relevância na explicação da variação na CPUE. Esses eventos climáticos globais influenciam diretamente as condições oceânicas e atmosféricas, afetando a disponibilidade de alimento, padrões de migração e comportamento dos peixes, e, consequentemente, a eficiência da pesca.

Ao adotar um modelo que considera esses fatores inter-relacionados, somos capazes de melhorar nossa compreensão das complexas dinâmicas da atividade pesqueira e, por conseguinte, auxiliar na formulação de estratégias de manejo mais eficazes e sustentáveis. No entanto, é importante ressaltar que qualquer modelo de previsão está sujeito a limitações e incertezas inerentes à complexidade dos sistemas naturais.

Apêndice

Códigos utilizados para obter o modelo de regressão:

```
library(papeR)

dados2 <- dados2 |> select(-c(cpue))

modelo <- stats::lm(cpue2 ~ ., data=dados2)
summary(modelo)
```

```
##
## Call:
## stats::lm(formula = cpue2 ~ ., data = dados2)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.27187 -0.34282  0.01668  0.44013  1.25623
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -48.667901  78.102292  -0.623  0.53416
## frotaUbatuba  -0.204053   0.132931  -1.535  0.12691
## ano            0.026670   0.039271   0.679  0.49811
## trimestre     -0.007776   0.047373  -0.164  0.86985
## latitude       0.209544   0.072329   2.897  0.00434 **
## longitude     -0.112947   0.076929  -1.468  0.14417
## fenomenonina   0.275510   0.177370   1.553  0.12249
## fenomenonino   0.440035   0.170317   2.584  0.01074 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5811 on 148 degrees of freedom
## Multiple R-squared:  0.2057, Adjusted R-squared:  0.1682
## F-statistic: 5.476 on 7 and 148 DF,  p-value: 1.302e-05
```

```
opt_model_step_aic<- stepAIC(modelo, direction="both")
```

```
## Start:  AIC=-161.58
## cpue2 ~ frota + ano + trimestre + latitude + longitude + fenomeno
##
##             Df Sum of Sq    RSS     AIC
## - trimestre  1   0.00910 49.983 -163.56
## - ano        1   0.15574 50.130 -163.10
## <none>                   49.974 -161.58
## - longitude  1   0.72786 50.702 -161.33
## - frota      1   0.79564 50.770 -161.12
## - fenomeno   2   2.38971 52.364 -158.30
## - latitude   1   2.83402 52.808 -154.98
##
## Step:  AIC=-163.55
## cpue2 ~ frota + ano + latitude + longitude + fenomeno
##
##             Df Sum of Sq    RSS     AIC
## - ano        1   0.16990 50.153 -165.03
## <none>                   49.983 -163.56
## - longitude  1   0.75262 50.736 -163.22
## - frota      1   0.79943 50.782 -163.08
## + trimestre  1   0.00910 49.974 -161.58
## - fenomeno   2   2.45062 52.434 -160.09
## - latitude   1   2.82572 52.809 -156.98
##
## Step:  AIC=-165.03
## cpue2 ~ frota + latitude + longitude + fenomeno
##
##             Df Sum of Sq    RSS     AIC
## <none>                   50.153 -165.03
## - longitude  1   0.65682 50.810 -165.00
```

```
## - frota      1   0.78379 50.937 -164.61
## + ano        1   0.16990 49.983 -163.56
## + trimestre  1   0.02326 50.130 -163.10
## - fenomeno   2   2.96926 53.122 -160.05
## - latitude   1   2.67032 52.823 -158.93
```

```
summary(opt_model_step_aic)
```

```
##
## Call:
## stats::lm(formula = cpue2 ~ frota + latitude + longitude + fenomeno,
##     data = dados2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32339 -0.33798  0.03405  0.41746  1.25574
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.43907    2.30868   1.923  0.05640 .
## frotaUbatuba -0.20244    0.13222  -1.531  0.12786
## latitude      0.19974    0.07068   2.826  0.00536 **
## longitude    -0.10525    0.07509  -1.402  0.16310
## fenomenonina  0.33883    0.14906   2.273  0.02444 *
## fenomenonino  0.47914    0.16081   2.980  0.00337 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5782 on 150 degrees of freedom
## Multiple R-squared:  0.2029, Adjusted R-squared:  0.1763
## F-statistic: 7.635 on 5 and 150 DF,  p-value: 2.034e-06
```